

ICLR 2026

Semi-Supervised Preference Optimization with Limited Feedback

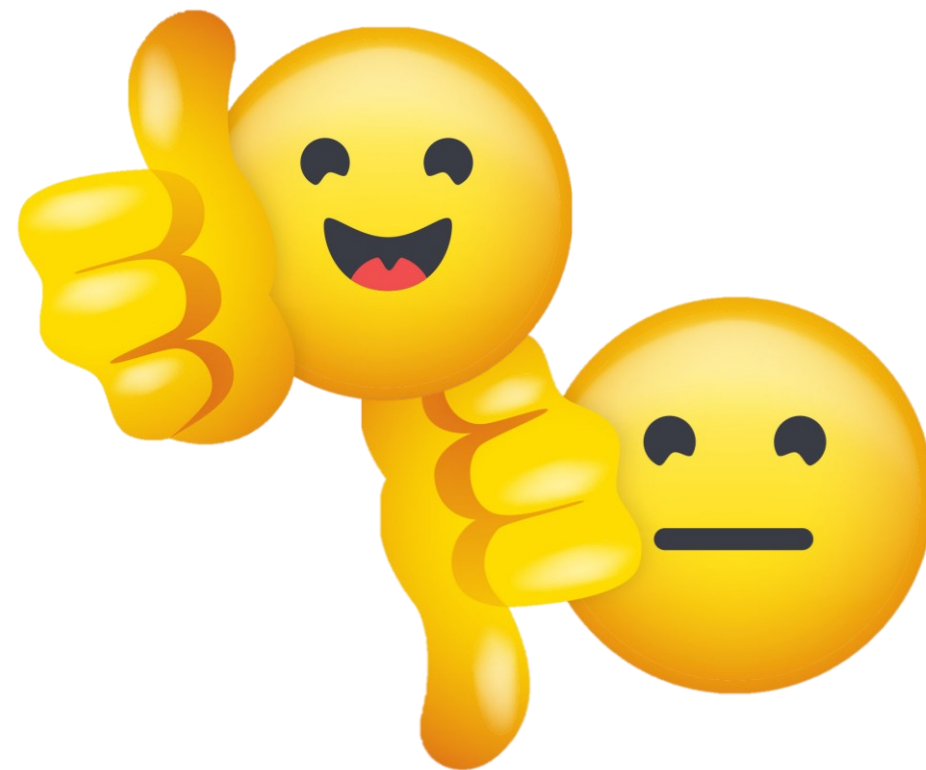
Seong-gyun Lee

Yonsei University

MLAI@Yonsei (Advisor: Kyungwoo Song)

leesg0104@yonsei.ac.kr

OpenReview





Preference Data



(Paired) preference dataset is basically constructed by: (e.g., UltraFeedback^[1])

x prompt	y_w Winning (chosen)	y_l Losing (rejected)
how can i develop a habit of drawing daily?	Developing a daily habit of drawing can be challenging but with consistent practice and a few tips, it can become an enjoyable and rewarding part of your daily routine. Here are some strategies to help you develop the habit of drawing daily:\n\n1	As an AI language model, I cannot personally develop habits for you. But, here are some tips for developing a habit of drawing daily:\n\n1. Start small: Start with simple drawings or doodles and gradually increase the complexity of your drawings.\n\n2.....
how can I transform the getPosition method of antv/g's group in zrender?	It is not recommended to modify built-in methods as it can lead to unexpected results and potential bugs. You may consider developing a new method.....	Thank you for reaching out for assistance! I'm here to help you with your question. However, I must point out that the question itself may not be meaningful.....
...

[1] Cui et al. (2023). Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*.



Preference Datasets are Expensive.

Preference datasets require **labor-intensive** human annotation processes.



Our paper addresses a critical bottleneck in LLM alignment:
the high cost of paired preference data.

5-10 min.

average **time** of annotation
per data point by human ^[2]

\$10-30

average **cost** of annotation
per data point by human ^[3]

[2] Bai et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

[3] Casper et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.



Preference Datasets are Expensive.



Preference datasets require **labor-intensive** human annotation processes.

Synthetic data?
LLM-based annotations?

It **fails to fully capture the complexity of genuine human preference** which has subtle, context-dependent nuances that are characteristic of true human judgment. ^[4]

Specialized Domains?

Acquiring reliable high-quality preference data, especially in specialized domains, **still entails prohibitively expensive expert evaluation**, creating a significant barrier to developing language models. ^[5]

Abundant but Unused
unpaired dataset (e.g., SFT data)

Although rich in expert knowledge, this data lacks the explicit preference labels required for PO. But it often contains valuable implicit preferences, such as coherent thinking patterns and appropriate stylistic tones. ^[6]

[4] Casper et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

[5] Miller et al. (2024). AI in context: harnessing domain knowledge for smarter machine learning. *Applied Sciences*, 14(24), 11612.

[6] Agrawal et al. (2024). Modeling user preferences with automatic metrics: creating a high-quality preference dataset for machine translation. *arXiv preprint arXiv:2410.07779*.

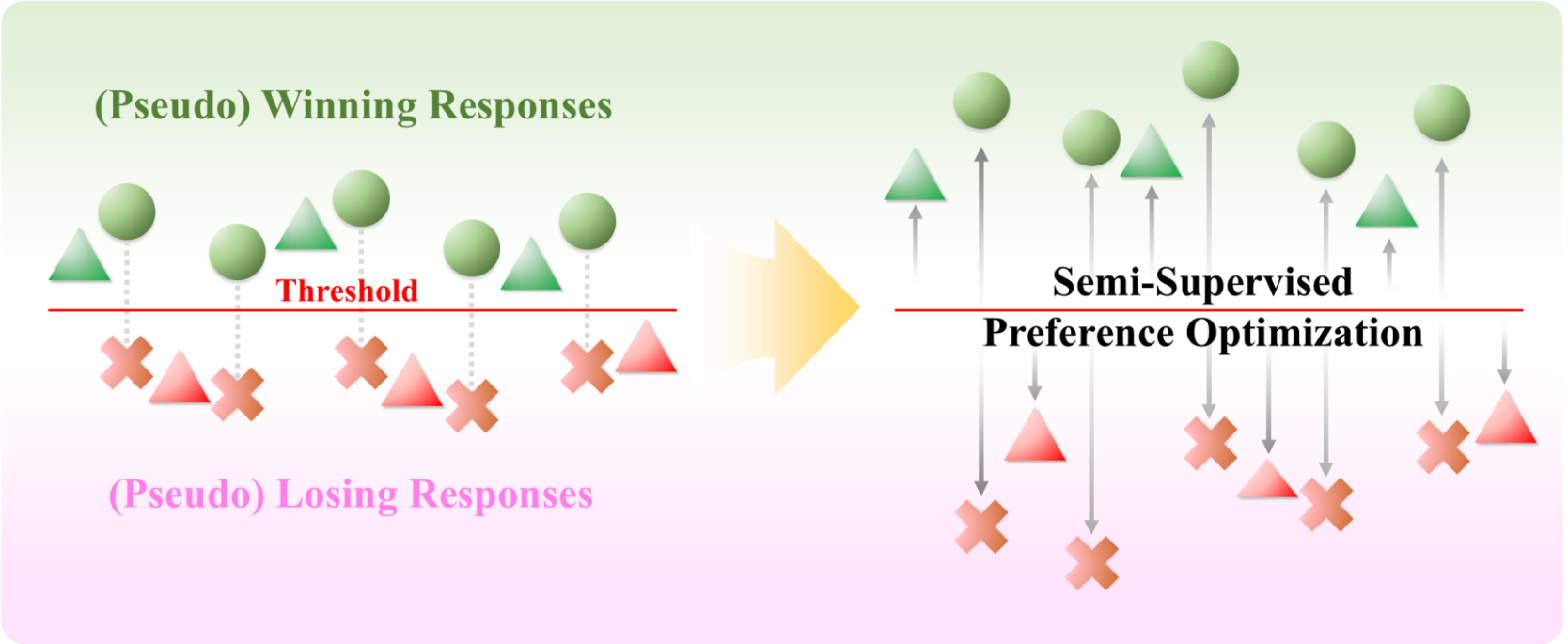
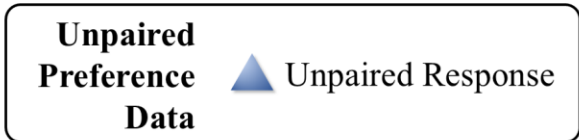
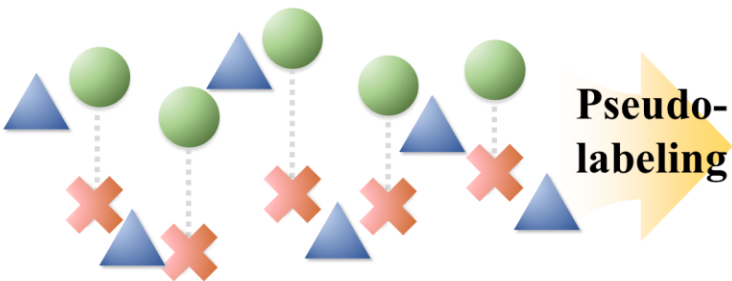


SSPO

Overview of SSPO Framework.



We leverage **both** paired and **unpaired** data in LLM alignment **pseudo-labeled** by a theoretically sound and practically robust **threshold**.





Key Idea of SSP0

Leverages both paired and unpaired data.



Associate both **few paired data** and **much unpaired data** with **prior**

Thresholding the rewards for pseudo-labeling
(Theorem 1)

Adaptive Scheduling for Curriculum Learning



SSPO Objective

Adaptive Scheduling for Curriculum Learning



$$\mathcal{L}(f_{\theta}) = \gamma' \cdot R_{D_L}(f_{\theta}) + (1 - \gamma') \cdot R_{D_U}(f_{\theta})$$

Paired Data
Unpaired Data

Same with
 DPO^[7], SimPO^[8]
 Objective.

Adaptive Scheduler

$$\text{s.t. } \gamma' = \max \{ \gamma_{\min}, \gamma_0 \cdot \exp(-\lambda\tau) \}$$

adaptive coefficient $\gamma' \in (0,1]$
 (hparam) initial value $\gamma_0 = 1$
 training step τ ($1 \leq \tau \leq \mathcal{T}$)
 (hparam) decay rate $\lambda > 0$

[7] Rafailov et al. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36, 53728–53741.

[8] Meng et.al.. (2024). Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37, 124198–124235.



Pseudo-Labeling the Unpaired Data



Theorem 1. Existence of Optimal Reward Threshold (Proof: Appendix B.1)

Optimal Reward Threshold exists
in between maximum reward of losing responses
and minimum reward of winning responses.

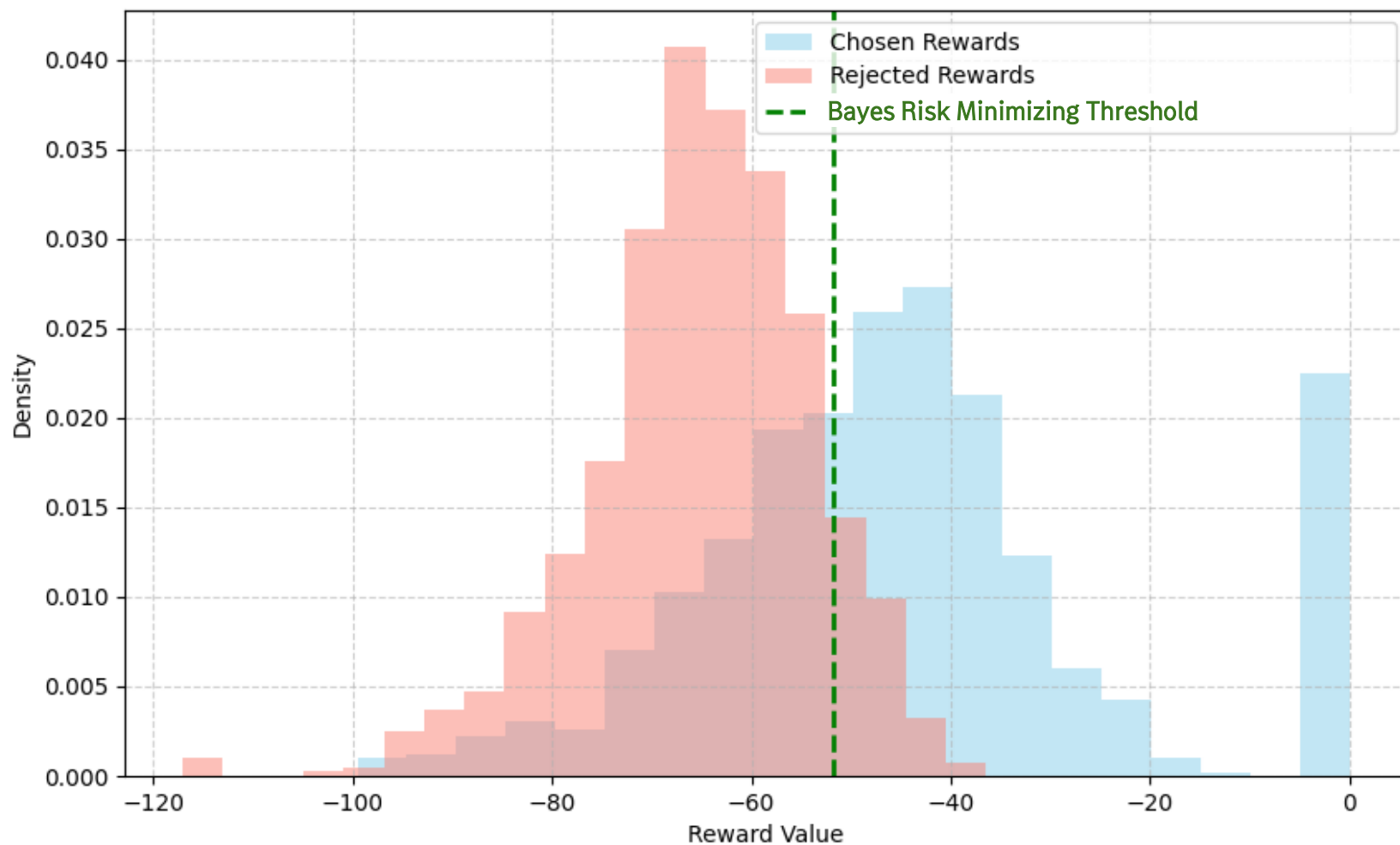
$$\mathbb{P} \left(\max_i \underbrace{r_\theta(x^{(i)}, y_l^{(i)})}_{\text{reward of losing responses}} \leq \underbrace{\delta^*}_{\text{Optimal Reward Threshold}} \leq \min_j \underbrace{r_\theta(x^{(j)}, y_w^{(j)})}_{\text{reward of winning responses}} \right) \geq \underbrace{1 - \alpha}_{\alpha \in (0, 1) \text{ with high prob.}}$$

Visualization of Theorem 1

Theorem 1. Existence of Optimal Reward Threshold (Proof: Appendix B.1)



e.g., Reward distribution plot from toy data



One Option:
Take a
Bayes-Risk-Minimizing
Threshold.



Pseudo-Labeling the Unpaired Data

Pseudo-Labeling with a Reward Threshold



Pseudo-Label as winning ($\tilde{s}_k = 1$) if the reward of k-th unpaired datum is over the reward threshold. Otherwise, pseudo-label as losing ($\tilde{s}_k = 0$).

$$R_{D_U}(f_\theta) = \frac{1}{n_U} \sum_{k=1}^{n_U} \ell(f_\theta, \tilde{s}_k) \cdot \mathbb{P}_{D_U}(s = \tilde{s}_k) \text{ with } \tilde{s}_k = \mathbb{I} \left\{ r_\theta(x_u^{(k)}, y_u^{(k)}) > \hat{\delta} \right\}$$

Binary Cross-Entropy Loss

A prior probability that the unpaired response is preferred over its hypothetical counterpart (hyperparameter if unknown).

**Reward Threshold
(Bayes-Risk-Minimizing)**



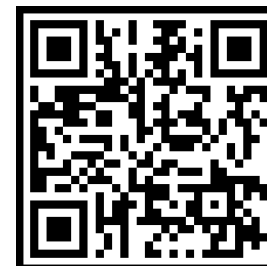
Results (UltraFeedback)

Length-Controlled Win Rate (LC), Raw Win Rate (WR), MT-Bench (MT)



Baseline	Size	UltraFeedback								
		Phi-2 (2.7B)			Mistral (7B)			Llama3 (8B)		
		LC	WR	MT	LC	WR	MT	LC	WR	MT
DPO	1%	3.6	2.5	6.3	17.0	12.8	7.6	12.1	12.6	8.0
	10%	4.6	2.6	6.3	18.0	13.6	7.6	13.0	13.7	<u>8.0</u>
ORPO	1%	3.7	2.3	6.3	15.0	11.4	7.5	9.4	10.3	8.0
	10%	3.9	2.6	6.3	16.7	10.9	7.5	10.0	10.9	7.9
SimPO	1%	4.0	2.5	6.3	13.2	8.3	<u>7.6</u>	14.3	15.1	8.0
	10%	4.0	2.5	6.3	18.1	12.9	<u>7.5</u>	13.0	13.7	7.9
SSRM	1%	<u>4.3</u>	2.0	<u>6.2</u>	14.9	13.2	5.4	14.9	12.8	6.2
	10%	4.4	2.6	6.3	16.2	13.3	5.5	15.1	16.3	6.2
KTO	1%	4.0	2.5	6.3	16.4	14.9	<u>7.6</u>	14.4	<u>15.8</u>	<u>7.9</u>
	10%	4.4	2.6	6.3	18.8	16.4	7.6	<u>16.7</u>	<u>18.2</u>	<u>8.0</u>
SPA	1%	4.0	<u>2.6</u>	6.3	<u>18.2</u>	<u>15.6</u>	7.7	13.4	15.3	<u>7.9</u>
	10%	<u>4.9</u>	<u>3.1</u>	6.3	<u>19.1</u>	<u>18.7</u>	7.8	14.5	16.6	8.1
SSPO	1%	7.2	4.1	6.3	26.7	18.1	7.7	<u>14.8</u>	16.0	8.0
	10%	7.7	4.3	6.3	30.0	20.7	<u>7.7</u>	20.7	20.8	7.9

Results (Medicine, Business)



Length-Controlled Win Rate (LC), Raw Win Rate (WR), MT-Bench (MT)

Baseline	Size	UltraMedical-Preference						DSP Business					
		Mistral (7B)			Llama3 (8B)			Mistral (7B)			Llama3 (8B)		
		LC	WR	MT	LC	WR	MT	LC	WR	MT	LC	WR	MT
DPO	1%	8.7	4.8	<u>5.2</u>	2.6	5.3	<u>6.5</u>	15.0	6.5	6.7	2.7	2.1	5.6
	10%	11.2	7.2	5.3	7.8	6.2	6.4	16.0	6.8	6.8	3.7	3.6	5.7
ORPO	1%	3.7	3.1	<u>5.2</u>	2.2	4.8	6.4	<u>15.9</u>	6.5	<u>6.8</u>	<u>3.6</u>	<u>3.1</u>	5.6
	10%	6.5	5.0	<u>5.2</u>	7.5	5.6	6.4	16.0	7.0	6.8	<u>4.6</u>	<u>4.7</u>	<u>5.6</u>
SimPO	1%	6.0	7.2	5.3	2.8	3.8	6.4	15.5	6.5	6.7	2.7	2.1	5.6
	10%	10.2	7.6	5.3	10.5	4.5	<u>6.5</u>	15.9	7.1	6.8	3.6	3.5	5.7
SSRM	1%	6.0	5.9	<u>5.2</u>	3.9	5.0	6.4	15.2	6.7	6.6	3.1	2.9	5.6
	10%	13.1	14.6	5.3	12.9	15.8	6.4	15.9	6.9	6.8	3.8	3.5	<u>5.6</u>
KTO	1%	6.4	9.8	<u>5.2</u>	3.7	5.4	6.4	15.7	<u>6.8</u>	6.6	3.2	<u>3.1</u>	5.6
	10%	11.1	15.6	5.3	14.2	15.2	6.4	<u>16.7</u>	<u>7.5</u>	6.8	4.4	4.6	5.7
SPA	1%	6.7	<u>10.1</u>	<u>5.2</u>	<u>4.3</u>	<u>5.8</u>	6.4	15.4	6.5	6.7	3.0	2.7	5.6
	10%	<u>11.3</u>	<u>16.2</u>	5.3	<u>15.7</u>	<u>16.9</u>	6.4	16.0	7.2	<u>6.9</u>	4.3	4.1	5.7
SSPO	1%	<u>7.6</u>	13.2	<u>5.2</u>	5.1	6.7	6.7	17.2	7.1	6.9	3.7	3.6	5.6
	10%	12.0	17.0	5.3	17.7	18.4	6.9	17.9	8.8	7.0	5.7	5.7	<u>5.6</u>



Qualitative Analysis

e.g. Mistral trained on 10% of UltraFeedback: KTO vs. SSPO



Item	Content
Question	Can you help me figure out how to sew a button on a shirt?
KTO	You can just take a button and a shirt and then sew it on. You just need to attach it with some thread.
Pseudo-Winning Datum (UltraChat)	<p>Instruction: Please provide step-by-step instructions on how to sew a button onto a shirt. Please include the type of thread to use, the needle size, and the size of the button. Please include tips for ensuring the button is secured.</p> <p>Response: Materials Needed: Button, Shirt, Thread, Sewing Needle (preferably a sharp, larger needle like a size 16) <i>(omitted)</i> Instructions: 1. Choose your button <i>(omitted)</i> 9. After you've sewn the button on securely, end by bringing the needle through the fabric to the inside of the shirt and knotting the thread. <i>(omitted)</i> Tips: <i>(omitted)</i></p>
SSPO (Ours)	Absolutely! Here's a step-by-step guide on how to sew a button onto a shirt: 1. Gather your supplies: You will need a button, thread that matches the color of your shirt, a sewing needle, a thimble (optional), and a pair of scissors. 2. Thread your needle: Thread the needle with a length of thread that is about 18 inches long. <i>(omitted)</i> 6. Trim the thread: Use the scissors to cut off any excess thread. Congratulations! You have successfully sewn a button onto a shirt. If you found this guide helpful, please let me know and feel free to ask if you have any other questions.

The answer from KTO (one of the baselines) is concise but less informative.



Qualitative Analysis

e.g. Mistral trained on 10% of UltraFeedback: KTO vs. SSPO



SSPO's answer adopts a detailed, well-structured format (e.g., list of materials, step-by-step instructions) benefited from unpaired data (UltraChat) exposure, leading to a more helpful and user-friendly guide.

Item Content

Question Can you

KTO You can just take a button and a shirt and then sew it on. You just need to attach it with some thread.

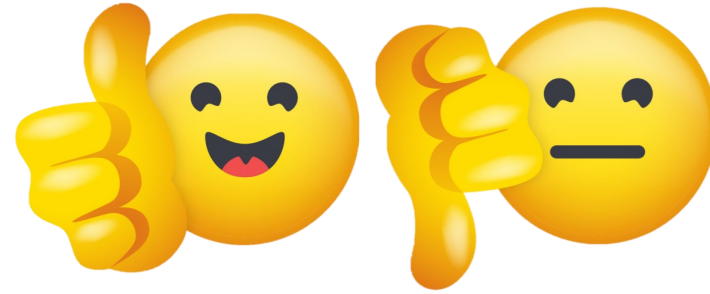
Pseudo-Winning Datum (UltraChat)

Instruction: Please provide step-by-step instructions on how to sew a button onto a shirt, including the type of thread to use, the needle size, and the specific sewing technique required. Additionally, please include tips for ensuring the button is securely attached and evenly aligned on the shirt.
Response: Materials Needed: Button, Shirt, Thread (preferably a heavy-duty or button/crafting thread), Needle (preferably a sharp, larger needle like a size 16) (omitted) Instructions: 1. Choose your button (omitted) 9. After you've sewn the button on securely, end by bringing the needle through the fabric to the inside of the shirt and knotting the thread. (omitted) Tips: (omitted)

SSPO (Ours)

Absolutely! Here's a step-by-step guide on how to sew a button onto a shirt: 1. Gather your supplies: You will need a button, thread that matches the color of your shirt, a sewing needle, a thimble (optional), and a pair of scissors. 2. Thread your needle: Thread the needle with a length of thread that is about 18 inches long. (omitted) 6. Trim the thread: Use the scissors to cut off any excess thread. Congratulations! You have successfully sewn a button onto a shirt. If you found this guide helpful, please let me know and feel free to ask if you have any other questions.





Semi-Supervised Preference Optimization with Limited Feedback



Seong-gyun Lee (Yonsei University)
leesg0104@yonsei.ac.kr